Rejoinder to "A Tuning-free Robust and Efficient Approach to High-dimensional Regression"

Lan Wang, Bo Peng, Jelena Bradic, Runze Li and Yunan Wu

We heartily thank the Editors, Professors Regina Liu and Hongyu Zhao, for featuring this paper and organizing stimulating discussions. We are grateful for the feedback on our work from the three distinguished discussants: Professors Jianqing Fan, Po-Ling Loh and Ali Shaojie. The discussants provide novel methods for inference, offer new applications such as graphical models and factor models, and highlight the possible impact of robust procedures in new domains. Their discussions have pushed forward robust high-dimensional statistics in disparate directions. These in-depth discussions with new contributions would easily qualify on their own as independent papers in the field of robust high-dimensional statistics. We sincerely thank the discussants for their time and effort in providing insightful comments and for their generosity in sharing their new findings. In the following, we organize our rejoinder around the major themes in the discussions.

1 Tuning parameter selection

We first briefly review the complete pivotal property of Rank Lasso, which enables the use of a simulated tuning parameter. Recall that the rank loss function is $Q_n(\boldsymbol{\beta}) = \big[n(n-1)\big]^{-1} \sum\sum_{i\neq j} \big|(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - (Y_j - \mathbf{x}_j^T\boldsymbol{\beta})\big|$. The subgradient of $Q_n(\boldsymbol{\beta})$, evaluated at the true

parameter value $\boldsymbol{\beta}_0$, is

$$\mathbf{S}_n = \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \;=\; -2\big[n(n-1)\big]^{-1}\sum_{j=1}^{n}\mathbf{x}_j\Big(\sum_{i=1,i\neq j}^{n}\text{sign}(\epsilon_j-\epsilon_i)\Big).$$

where $\text{sign}(t) = 1$ if $t > 0$, $= -1$ if $t < 0$, and $= 0$ if $t = 0$. Observe that $\xi_j = \sum_{i=1,i\neq j}^{n}\text{sign}(\epsilon_j-\epsilon_i) = 2\text{rank}(\epsilon_j)-(n+1)$, where $\text{rank}(\epsilon_j)$ is the rank of ϵ_j among $\{\epsilon_1,\ldots,\epsilon_n\}$. Since $(\text{rank}(\epsilon_1),\ldots,\text{rank}(\epsilon_n))^T$ follows uniform distribution on the set of permutations of integers $\{1,2,\ldots,n\}$, the distribution of $\mathbf{S}_n$ depends on neither $\boldsymbol{\beta}_0$ nor $\boldsymbol{\epsilon}$. Specifically, if we denote with $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_n)^T$, then given $\mathbf{X}$, the conditional distribution of $\mathbf{S}_n = -2\big[n(n-1)\big]^{-1}\mathbf{X}^T\boldsymbol{\xi}$, is both a completely known distribution and is independent of the random error distribution.

The above complete pivotal property enables us to select tuning parameter without needing to pre-estimate any unknown population quantity (such as $\boldsymbol{\beta}_0$) and automatically adjusts to both the random error distribution and the design matrix correlation structure. For any given c and α_0, we take λ equal to

$$\lambda^* = cG^{-1}_{||\mathbf{S}_n||_\infty}(1-\alpha_0)$$

where $G^{-1}_{||\mathbf{S}_n||_\infty}(1-\alpha_0)$ denotes the $(1-\alpha_0)$-quantile of the distribution of $||\mathbf{S}_n||_\infty$ (see Algorithm 1). In contrast, square-root Lasso has a partial pivotal property. This can be seen by observing that the gradient of its loss function, evaluated at $\boldsymbol{\beta}_0$, has the form $\big(\sum_{i=1}^{n}\epsilon_i^2\big)^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i$, which itself does not depend on the standard deviation σ, of the random error ϵ, but still depends on the other aspects of the error distribution (such are higher moments, for example). The method of Sun and Zhang (2012) is based on the normal error assumption of ϵ.

Algorithm 1 Simulated λ ($\mathbf{X}$, $\alpha = 0.1$, $c = 1.01$, times$= 1000$)

1: **for** k in 1:times **do**
2: Generate random perturbation for $1:n$, denoted as $\boldsymbol{\tau}$.
3: $\boldsymbol{\epsilon} = 2 * \boldsymbol{\tau} - (n+1)$.
4: $S[k] = c * \left|\left| \frac{-2}{n(n-1)} \sum_{i=1}^{n} \epsilon_i \mathbf{x}_i \right|\right|_{\infty}$.
5: **end for**
6: Output $\lambda = \text{Quantile}(\mathbf{S}, 1-\alpha)$.

In their discussions, Fan, Ma and Wang (FMW hereafter) propose a tuning parameter selector by using the asymptotic distribution of

$$\mathbf{S}_n = -\frac{2}{\sqrt{3}} \frac{n+1}{\sqrt{n}(n-1)} \frac{1}{\sqrt{n}} \mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. They further suggest a multiplier bootstrap appraoch to estimating the tuning parameter, which also precludes the dependence on $\boldsymbol{\beta}_0$ and the distribution of ϵ. Observe that the suggested approach necessitates asymptotic distributional approximations. Therefore some extra conditions may be needed to ensure the validity of the asymptotic approximation. In contrast, our approach does not rely on any asymptotic approximation and directly obtains the finite sample distribution of $\mathbf{S}_n$ for any n and p. Our simulations (Figure R.1) suggests that the tuning parameter based on asymptotic theory tends to be larger than the one based on finite-sample simulation of Algorithm 1 proposed in our paper.

2 Subsampling Calculations of Rank Lasso

We agree with FMW that as the rank loss function has the form of U-statistics, the computation of Rank Lasso may not be scalable due to its complexity of $O(n^2)$. They further proposed a subsampled version of the Rank Lasso estimator and named it ANOPE (Average Non-Overlapping Pairwise difference Estimator). ANOPE with a small ($m = 5$) is suggested to provide a satisfactory tradeoff between the computational cost and the estimation accuracy of the resulting estimate.

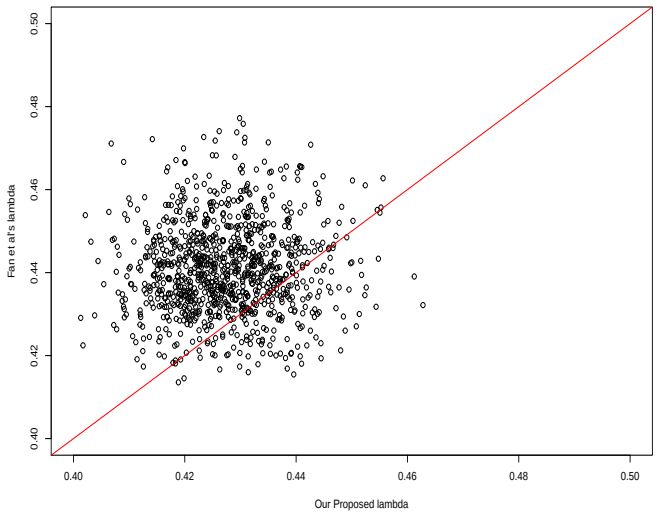

Figure R.1: Comparison of tuning parameters

Algorithm 2 Incomplete U-statistics algorithm ($\mathbf{X}$, $\mathbf{y}$, m)

1: Compute simulated λ given $\mathbf{X}$.
2: Compute $(\mathbf{x}_i - \mathbf{x}_j, y_i - y_j)$ for $1 \leq i < j \leq n$.
3: Generate random subsample $\mathcal{S} \subseteq \{(i,j) : 1 \leq i < j \leq n\}$ with $|\mathcal{S}| = mn$.
4: Compute the Rank Lasso estimator:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{mn} \Big\{ \sum_{(i,j)\in\mathcal{S}} |(y_i - y_j) - (\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\beta}| + \lambda ||\boldsymbol{\beta}||_1 \Big\}.$$

5: Output $\widehat{\boldsymbol{\beta}}$.

In our paper, we have focused on the on large p small n setting. For large n, we suggested using sub-sampling strategy (Algorithm 2). Our strategy shares the similarity with ANOPE that both algorithms are based on (mn)-pairs of the data $\{(i,j) : 1 \leq i < j \leq n\}$, albeit somewhat different random mechanisms are applied to select those pairs. A caveat is that the selection of the tuning parameter λ plays an important role in the performance of both algorithms.

We investigated the same simulation example as described in Section 2 of FMW. Figure R.2 compares Rank Lasso based on the incomplete U-statistics algorithm (Algorithm 2), ANOPE and the estimator based on the full sample, where the simulated λ (as proposed

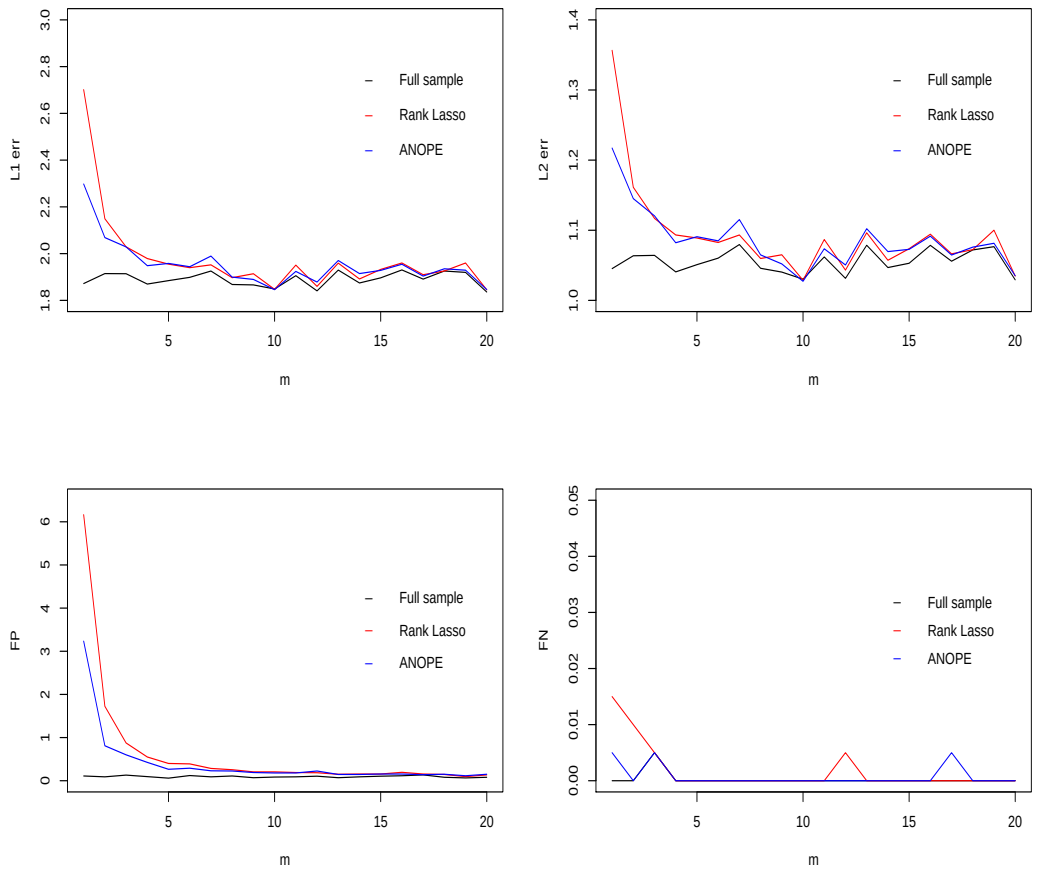

Figure R.2: Comparison of the U-statistic approximation: Full sample denotes a Rank-Loss computation based off of a complete sample, ANOPE is as proposed in Fan, Ma and Wang and Rank Lasso is as in Algorithm 2

.

in our paper) is applied to all three Algorithms. We report both estimation losses, through l_1 and l_2 estimation error, as well as variable selection performance, through False Positive (FP) and False Negative (FN) measures. We observe that when $m = 10$ or larger both the incomplete U-statistics algorithm and ANOPE perform similarly as the estimator based on the full sample. When m is smaller, both estimators perform notably suboptimal when compared with the full-sample estimator. Based on our observations, we would suggest to use either algorithm based on $10n$ or more pairs of the original data.

3 High-dimensional inference based on Rank Lasso

Statistical inference in high dimension is a very important and challenging problem. Both FMW and Li and Shojaie (LS hereafter) proposed new methods, although coming from two different perspectives, for high-dimensional robust inference by building off of Rank Lasso.

FMW provided a heuristic derivation of the debiased estimator of the Rank Lasso, based on the idea of inverting the first-order condition and extending the approach in Van de Geer et al. (2014). They also empirically demonstrated the validity of the debiased estimator.

LS empirically investigated and compared two alternative methods for the inference. They first considered a Wald-type inference procedure using a refitted approach. They propose to refit, without any penalty, the rank-based model consisting of only variables selected by rank SCAD; see Zhao et al. (2017). LS then studied a de-correlated score approach (one-step approach), motivated by Ning et al. (2017). An interesting finding of their simulation experiments is that for the refitted procedure, using Rank SCAD performs better than Rank Lasso; for the one-step procedure, using Rank SCAD as the initial estimator also leads to improved performance when compared with Rank Lasso as the initial estimator. This can perhaps be explained by a smaller finite-sample bias of the SCAD penalty.

We commend the efforts of FMW and LS in exploring this important research direction. Robust high-dimensional inference has not been discussed much at all in the current literature. Even in Gaussian models debiased inference suffers from tuning parameter selection, especially for the estimation of $\mathrm{Cov}^{-1}(\mathbf{x})$. We are happy to see their preliminary results suggest promising performance of inference based on Rank LASSO/SCAD in the high-dimensional setting.

This topic without doubt deserves a deeper study. Establishing a rigorous theory for any of the above three reference procedures is highly nontrivial due to the nonsmoothness of the rank loss function. To obtain CIs for β_j's: both methods require estimation of the scale parameter $\int f^2(u)\,du$, where $f(u)$ denotes the error density function. This is challenging in

the high-dimensional settings. Currently, both FMW and LS took this quantity as known in their numerical studies. Can we estimate the density of the random error without imposing restrictive modeling assumptions on the high-dimensional regression model? It is worth pointing out that, beyond the estimation of an error density at zero, it is unclear how to proceed with the estimation of the above functional.

4 Comparisons with other robust procedures

Both FMW and Loh brought historical perspectives into their discussions. Loh, in particular, raised several important new insights on connecting traditional robust statistics with modern high-dimensional data analysis.

As Loh pointed out, the proposed Rank Lasso can be understood as finding a regression parameter in high dimension that minimizes an L-estimate of the scale of the residuals. Thus, it is possible to extend other known robust measures in the classical robust statistics literature for the same purpose. This raises an interesting and important question whether certain estimator would be optimal among a class of estimators.

To answer this question, it is essential to first come up with appropriate measures of desirable statistical properties of a robust estimator in high dimensional setting. In the classical setting, properties such as estimation efficiency, high breakdown properties have played important roles. Meaningful extension of these classical concepts to high dimensional setting is not entirely straightforward. Loh asked whether optimality results could be proved in terms of the variance of the estimators in finite samples. Any new theory in this direction would be important since finite-sample error bounds have been the focus in the current high-dimensional regression literature.

It is worth emphasizing that robust estimation in high dimension necessitates an evaluation of its performance from multiple aspects. Despite other choices of robust loss function, we believe rank loss (based on Wilcoxon scores) has the unique advantage of achieving an

appealing trade-off among robustness, estimation efficiency and computational convenience. We view this new approach as a useful complement to Lasso, and not as a replacement.

5 Extensions and future research directions

All three discussants have discussed interesting areas to which Rank Lasso can be extended. FMW considered a useful extension to a factor-adjusted regression model to account for strongly dependent covariates, and provided promising numerical results. They also suggested the importance of extension to heteroscedastic regression and beyond linear models. We would like to add that their extension could provide some new insight into causal inference where factor models are often a powerful tool for removing the confounding effects.

LS provided a valuable and detailed analysis of application of Rank Lasso and Rank SCAD to graphical modeling. In their numerical studies, it was observed that the tuning parameter selection can be quite difficult (more so than for linear regression models). For popular existing approaches such as glasso and npn, BIC for example often gives an empty graph. In contrast, rank Lasso and rank SCAD provided promising results even when the linear model is misspecified, suggesting that tuning-free rank Lasso has broader applications. We consider robust graphical modeling and robust precision matrix estimation to be very important research areas. There has been a recent stream of interesting work on robust estimation of high-dimensional precision matrices, see Avella-Medina et al. (2018), Loh and Tan (2018), Goes et al. (2020), among others. Related to that is a study of the model misspecification and the possible stability property of rank Lasso type methods.

Loh suggested analyzing estimators that are robust to adversarial perturbations, a topic that is of particular interest to the computer science area, see Duchi and Namkoong (2020); Carmon et al. (2019), among others. We have performed a small Monte-Carlo experiment to examine the performance of rank-based methods at the presence of contamination in the predictors and random error. We generate $\mathbf{X}$ and $\boldsymbol{\beta}_0$ as in Example 3 of the main

Table R.1: Performance of different methods with perturbed $\mathbf{X}$

Method	L1 error	L2 error	ME	FP	FN
Lasso	1.97 (0.07)	0.82 (0.03)	0.92 (0.05)	8.66 (0.38)	0 (0)
Lasso-1se	2.71 (0.08)	1.55 (0.04)	4.24 (0.20)	0 (0)	0 (0)
$\sqrt{\text{Lasso}}$	1.65 (0.06)	0.80 (0.03)	0.98 (0.05)	4.11 (0.18)	0 (0)
SCAD	1.42 (0.06)	0.84 (0.04)	0.77 (0.06)	0 (0)	0 (0)
Huber Lasso	1.83 (0.03)	1.03 (0.02)	1.85 (0.07)	0 (0)	0 (0)
Rank Lasso	0.39 (0.01)	0.22 (0.01)	0.07 (0.00)	0 (0)	0 (0)
Rank SCAD	0.26 (0.01)	0.19 (0.01)	0.04 (0.00)	0 (0)	0 (0)

Note: Lasso uses λ corresponding to the minimum of the cross-validation error, Lasso-1se is the cross-validated Lasso with λ selected using the one standard error rule.

paper, where X has an AR(1) correlation matrix with auto-correlation coefficient 0.5, and $\epsilon \sim 0.95N(0,1) + 0.05N(0,10^2)$. Let $\mathbf{X}$ be contaminated by the small error: $\mathbf{Z} = \mathbf{X} + \mathbf{W}$, where $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n\times p}$, $w_{ij} \sim \text{Unif}(-0.1, 0.1)$ are i.i.d. random variables. Then we estimate $\boldsymbol{\beta}_0$ based on $(\mathbf{Z}, \mathbf{y})$. We observe that Rank Lasso performs significantly better than Lasso. Rank SCAD is observed to have the best overall performance.

It is worth mentioning that Loh (2017) studied a class of generalized M-estimators using Mallows, Hill–Ryan and Schweppe type weight functions. She established a rigorous theory in the high dimensional setting and numerically demonstrated robust performance of this class of estimators to contamination in the predictors and/or the response variables. Rank Lasso and Rank SCAD proposed in this paper is not specifically adversarial perturbations of the covariates. A possible generalization is to incorporate similar weights as those in Loh (2017).

As seen above, there are ample opportunities to explore contrasting and overarching robustness properties of the proposed method. All have demonstrated broad practical relevance and deserve further in-depth studies. It is unclear how to quantify the tradeoffs between robustness and efficiency among all or some of the above discussed robustness quantifications. All discussants called for more methodological developments in addressing a number of robustness questions. We wholeheartedly agree with our discussants and hope that this article

and its discussions would stimulate further growth in that direction.

References

Avella-Medina, M., Battey, H. S., Fan, J., and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. (2019). Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11192–11203.

Duchi, J. and Namkoong, H. (2020). Learning models with uniform performance via distributionally robust optimization. *to appear Annals of Statistics*.

Goes, J., Lerman, G., Nadler, B., et al. (2020). Robust sparse covariance estimation by thresholding Tyler's m-estimator. *The Annals of Statistics*, 48(1):86–110.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896.

Loh, P.-L. and Tan, X. L. (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, 12(1):1429–1467.

Ning, Y., Liu, H., et al. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202.

Zhao, S., Shojaie, A., and Witten, D. (2017). In defense of the indefensible: A very naive approach to high-dimensional inference. *arXiv preprint arXiv:1705.05543*.